DOCUMENT RESUME

ED 286 173 CS 210 513

AUTHOR Kang, Namjun
TITLE Alternative Methods for Calculating Intercoder Reliability in Content Analysis: Kappa, Weighted Kappa and Agreement Charts Procedures.
PUB DATE Aug 87
NOTE 22p.; Paper presented at the Annual Meeting of the Association for Education in Journalism and Mass Communication (70th, San Antonio, TX, August 1-4, 1987).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Content Analysis; *Interrater Reliability; *Measurement Techniques; Media Research; *Reliability; *Research Methodology; Theories; Theory Practice Relationship
IDENTIFIERS Coding; Empirical Research; Kappa Coefficient; Weighted Data

ABSTRACT

If content analysis is to satisfy the requirement of objectivity, measures and procedures must be reliable. Reliability is usually measured by the proportion of agreement of all categories identically coded by different coders. For such data to be empirically meaningful, a high degree of inter-coder reliability must be demonstrated. Researchers in various disciplines such as psychiatry, biometrics, epidemiology, and medical diagnostics have employed different procedures for their own specific problems; however, researchers in communication studies have not used parallel developments in these areas. The most widely used intercoder reliability measures in communication studies--Scott's Pi and the "index of crude agreement"--are based on questionable marginal homogeneity assumptions. Without the homogeneity assumption, J.A. Cohen proposed kappa statistic, a coefficient similar to Scott's Pi. The use of kappa and weighted kappa is restricted to the cases where the number of coders is two, but later researchers extended the concept to more than two coders. S.I. Bangdiwala and Hope Bryan have developed an "agreement chart" method that can help investigators pinpoint the source of poor intercoder agreement, and in which the use of graphics aids in grasping and remembering relationships. An analysis of fictitious data further illustrates Bangdiwala and Bryan's methodology. Many of the reliability measures may be executed by using a hand calculator (for items with only three categories) or "canned" computer programs. (Charts and tables of data are included, and figures and references are attached.) (NKA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286173

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ALTERNATIVE METHODS FOR CALCULATING INTERCODER RELIABILITY IN CONTENT ANALYSIS

-- Kappa, Weighted Kappa and Agreement Charts Procedures --

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Namjun Kang

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Namjun Kang

Rm. 24 NCC. I,
Newhouse School of Public Communication
215 University Place
Syracuse University
Syracuse New York, 13244
(315) 425-3163

CS210513

(Paper Submitted to AEJMC Research and Methodology Division)

Abstract

If content analysis is to satisfy the requirement of objectivity, measures must be reliable. Reliability measures is usually measured by the proportion of agreement between coders. The most widely used intercoder reliability measures in communication studies --Scott's Pi and the "index of crude agreement"-- are based on the questionable marginal homogeneity assumptions. Without the homogeneity assumption, Cohen proposed kappa statistic. The extensions of kappa -- weighted kappa, hierarchical kappa test, maximum kappa permitted by marginals-- are reviewed. The newly developed "agreement chart" method, which can help the investigator to pinpoint the source of poor intercoder agreement, is also examined. The apllication of these methods is provided.

I. Introduction

If content analysis is to satisfy the requirement of objectivity, which is a *sine qua non* of any scientific inquiry, measures and procedures must be reliable; i. e., repeated measures with the same instrument on a given sample should yield similar results. Reliability is usually measured by the proportion of agreement of all categories identically coded by different coders. When agreement is due merely to chance, data may have little to do with the phenomena studied. For such data to be empirically meaningful, a high degree of inter-coder reliability must be demonstrated (Krippendorff, 1970). Since the construction of the data is usually beyond the investigator's control in content analysis, opportunities for enhancing reliability are generally limited to improving coders, categories, or both. This is the major reason that intercoder reliability, which depends on the skill of the coders, has been regarded as the adequate reliability concept in content analysis as compared with the test-retest reliability concept, which has been primarily concerned with the testing instrument .

The analysis of intercoder reliability has received attention in a wide range of research disciplines as reviewed by Landis and Koch (1975a,1975b). Researchers in various disciplines such as psychiatry, biometrics, epidemiology, and medical diagnostics have employed different procedures for their own specific problems; however researchers in communication studies have not utilized the parallel developments in these other areas. They have been the die-hard users of Scott's Pi coefficient proposed 30 years ago with somewhat questionable assumptions and the innocent users of the most naive intercoder reliability index -- the "index of crude agreement." Futhermore, there appeared to have been a general reluctance to recognize coder errors in content analysis, and even more hesitation to analyze this coder variability in some rigorous, quantitative manner. For example, among sixteen articles in volume 61(1984) of ***Journalism Quarterly*** coded in the subject index for that volume under the "content analysis" heading, only eight articles reported the intercoder reliability measures. Among the eight reported reliability measures, five were the "index of crude agreement," two belonged to Scott's Pi, and one was Pearson correlaton coefficient *r*.

One reason the communication researchers in content analysis have not paid much attention to the intercoder reliability measures may be that the conventional reliability measures are not as useful as they could be. The researchers measure the reliabilty in the hope that this measure will be able to reveal what is wrong in content analysis. When the

measure reveals poor agreement among coders, the traditional reliability measures like Scott's Pi and "index of crude agreement" cannot locate the source of poor agreements. Errors may result from defective category constructions (e.g., ambiguous categories), or they may originate in the coders. Like Pearson's product moment correlation in parametric statistics, the traditional reliabilty measures are directionless indices.

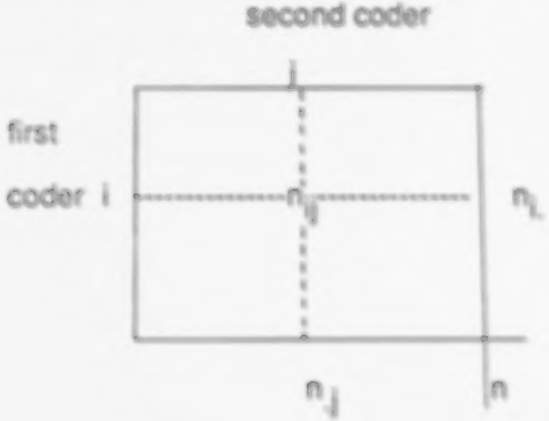
This paper will review the alternative reliability measures to Scott's Pi, the "index of crude agreement," and the way of presenting the intercoder reliability graphically to help the investigators to pinpoint the source of poor inter-coder agreement.

II. Different Measures of Inter-coder Reliability

As reviewed in Landis and Koch (1975a, 1975b), a wide variety of estimation and testing procedures have been recommended for the assessment of coder reliability. Among these various procedures, concern is limited to the following three methods, which have been widely used and known in communicatuions studies.

1) The Index of Crude Agreement

Data collected to assess the agreement between two independent coders can always be represented in a square matrix:



where n_{ij} is the number of observations that are assigned to category I by the first coder and to category J by the second coder, $n_{i.}$ is the total frequency with which the category I is used by the first coder, $n_{.j}$ is the total frequency with which the category J is used by the second coder, and n is the total number of observations recorded by both coders. The diagonal entries (n_{ii}) contains the incidents of agreement. If scales are categorical or nominal, any deviation from the diagonal consists of disagreement.

The most elementary index of agreement is based on the proportion of the subjects classified in the same category by the two coders and can be estimated by

$$P_O = \sum_i P_{ii} \qquad \text{---------- (1)}$$

where P_o is the "index of the crude agreement," and $P_{ii}=1/n\ (n_{ii})$. This index is the percentage of judgments on which coders agree, out of the total number of judgments.

This agreement index has been criticized, because it does not consider the extent of inter-coder agreement which may result from chance. This measure is biased in favor of items with a small number of categories. By chance alone, agreement should increase as the number of categories decreases. Thus, Scott (1955) derived a suitable correction for chance agreement to P_o.

2) Scott's Pi

Let PO denote the pecentage of judgments on which the two analysts agree when coding the same data independently, and let PE denote the percent agreement to be expected on the basis of chance. Then (PO-PE) represents the excess agreement beyond chance, and (1-PE) indicates the maximum possible excess agreement beyond chance. Then define

$$Pi = (PO-PE)/(1-PE) \qquad \text{---------- (2)}$$

as a standardized measure of agreement adjusting for the agreement due to chance. It can be roughly interpreted as the extent to which the coding reliability exceeds chance. In this context Pi has the following properties:

+1, if both coders are in complete agreement

0, if the level of agreement is equal to the expected agreement due to chance

<0 , if observed agreement is less than the expected agreement due to chance

In calculating Pi, the PE (expected agreement percentage) is computed on the basis of known population marginal probabilities, and the assumption that both coders have the identical marginal distributions of proportions, equal to the known population values. Thus PE are computed on a joint marginal distribution - that is, on the average frequencies in the categories, - then;

$$PE = 1/n^2 \sum_i ((n_{i.} + n_{.i})/2)^2 \qquad \text{---------- (3)}$$

Now, if equation (1) and (3) are inserted into equation (2), the Pi coefficient can be rewritten as follows;

$$Pi = \frac{n \sum_i n_{ii} - \sum_i ((n_{i.} + n_{.i})/2)^2}{n^2 - \sum_i ((n_{i.} + n_{.i})/2)^2}$$

$$= \frac{4n \sum_i n_{ii} - \sum_i (n_{i.} + n_{.i})^2}{4n^2 - \sum_i (n_{i.} + n_{.i})^2} \qquad \text{---------- (4)}$$

Krippendorff (1970, p. 147) proposed this formula as the efficient computation formula. Latter Krippendorff (1980) generalized this Pi statistic for cases having more than two coders and for ordinal and interval scales by assigning distance measurement weights. When the number of coders is exactly two, and the categories of the variables are unordered (nominal

scale), then Krippendorff's agreement coefficient equals Scott's Pi.

As already mentioned, Scott assumed for his Pi that the distribution of proportions over the categories for the population is known and is taken to be equal for the coders. Cohen (1960) agreed that the former assumption was reasonable for interviewer analysis in survey research, but the latter might be questioned in more general applications, since one source of disagreement between a pair of judges is precisely their proclivity to distribute their judgment over the categories. These different proclivities between coders are represented by their respective marginal distributions. Thus averaging the marginal distributional difference between coders -- the method used in calculating PE in Scott's Pi -- may wash out the intercoder differences that is the core substance the inter coder reliability measures are supposed to measure. Actually, Scott himself conceded that the assuption of marginal homogeneity was not necessary. " ... ... PE [expected frequency] can be calculated from the actual distributions of each pair of coders." But he added that "... ... [this] procedure is more complicated" (Scott, 1955 p. 324). This statement is , to some extent, exaggerated. Calculating PE based on actual distribution of each coder is no more complex than that of PE based on Scott's homogeneity assumption, especially when we use the computer program. In this context, Cohen (1960) proposed the kappa statistic, which takes into account the information contained in the marginal distributions of each coder.

3) Cohen's Kappa

A coefficient proposed by Cohen (1960) is quite similar to Scott's Pi. Here, PE is differently defined. In Cohen's agreement coefficients -- called kappa -- the expected frequencies are based on the observer's respective marginal distribution:

$$PE = 1/n^2 \sum_i n_{i.} n_{.i} \qquad \text{..... (5)}$$

When equation (1) and (5) are inserted into equation (2), then

$$\text{Kappa} = \frac{n \sum_i n_{ii} - \sum_i n_{i.} n_{.i}}{n^2 - \sum_i n_{i.} n_{.i}} \qquad \text{..... (6)}$$

If marginal frequencies are equal, that is, $n_{i.} = n_{.i}$ for all i, then equation (4) and (6)

are equal by:

$$\text{Pi=Kappa=} \frac{n\sum_i n_{ii} - \sum_i n_{i.}^2}{n^2 - \sum_i n_{i.}^2} \qquad \text{---------- (7)}$$

Thus, Pi and kappa merely vary with the way PE is computed. In particular, Scott's agreement measure Pi was derived under the simultaneous constraints of marginal homogeneity and coder independence, whereas cohen's kappa measure was developed using only the constraint of coder independence.

Since kappa is also based on the basic equation described in Eq. (2), the upper limit is +1 like Pi, occuring when (and only when) there is perfect agreement between the coders. And if the level of agreement is equal to the expected agreement due to chance, then kappa=0.

It is of some interest to estimate the maximum value of kappa set by the marginal distributions. It is

$$K_m = \frac{PO_m - PE}{1 - PE} \qquad \text{---------- (8)}$$

where PO_m is found by pairing the $P_{i.}$ ($P_{i.} = n_{i.}/n$) and $P_{.i}$ ($P_{.i} = n_{.i}/n$) values, selecting the smaller of each pair and summing the k values (Cohen, 1960). Thus, K_m is the maximum value of kappa allowed by the marginals, and $1-k_m$ represents the proportion of the possibilities of agreement (chance excluded) which cannot be achieved as a consequence of differing marginals. The latter value can be used as an indicator of the ambiguity of the category boundaries and may be reduced further by training of coders or sharpening the boundaries of categories.

In a latter paper, Cohen (1968) introduced a modified form of kappa which allows for scaled disagreement or partial credit. Since certain patterns of disagreement between the observers may be more important than others, he proposed using a set of weights $\{w_{ii'}\}$ for $i, i' = 1,2, \ldots k$, which reflect the contribution of each cell to the measure agreement. Accordingly, in order to measure agreement with respect to a specified set of weights $\{w_{ii'}\}$, Cohen defined a weighted

kappa measure which is estimated by

$$K_w = \frac{PO^{\bullet} - PE^{\bullet}}{1 - PE^{\bullet}} \quad \text{---------- (9)}$$

where

$$PO^{\bullet} = \sum_i \sum_{i'} W_{ii'} P_{ii'} \quad \text{and} \quad \text{---------- (10)}$$

$$PE^{\bullet} = \sum_i \sum_{i'} W_{ii'} P_{i.} P_{.i} \quad \text{---------- (11)}$$

In most cases, $o<=W_{ii'}<=1$ for all i, i', so $PO^{\bullet}$ can be interpreted as a weighted observed proportion of of agreement, and $PE^{\bullet}$ is the corresponding weighted proportion of agreement expected under the constraint of marginal distributional independence. Furthermore, by choosing the weights in (12),

$$W_{ii'} = \begin{cases} 1, & i = i' \\ 0, & i \neq i' \end{cases} \quad \text{---------- (12)}$$

K_w in (9) simplifies to kappa in (7). In other words, Kappa is a special case of K_w.

Formulas for the standard errors of kappa and K_w statistics have been given in the literatures (Cohen, 1960; Cohen, 1968; Everitt, 1968), but they were in error. Valid formulas for the approximate large sample variance are given in Fleiss, Cohen and Everitt (1969).

Agreement statistics such as K_w in (9) can also be generated for ordinal scale data by assigning appropriate weights to each of the off-diagonal cells to reflect the degree of disagreement. One such selection of weights recommended by Cichetti (1972) is given by

$$W_{ii'} = 1 - (|i - i'|)/(k-1) \quad \text{---------- (13)}$$

Table 1

Hierarchical Weights for Agreement Measures

Weights		W_1	W_2	W_3	W_4
Coder		B	B	B	B
	directional Scale	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
	1	1 0 0 0 0	1 1 0 0 0	0 0 0 0 0	1 1 0 0 0
	2	0 1 0 0 0	1 1 0 0 0	0 0 0 0 0	1 1 0 0 0
Coder A	3	0 0 1 0 0	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
	4	0 0 0 1 0	0 0 0 0 0	0 0 0 1 1	0 0 0 1 1
	5	0 0 0 0 1	0 0 0 0 0	0 0 0 1 1	0 0 0 1 1

We can use K_w to investigate the specific patterns of disagreement between the coders on category classification by selecting a hierarchy of weights. These kinds of weights successively combine the adjoining categories in order to create a potentially less stringent reliability measure (Landis and Koch, 1977b). The weights can be chosen so that the associated kappa measures indicate the increments in agreement which are induced by successively combine relevant categories of the response variable. For example, the four sets of weights in Table 1 can be used to investigate the sources of imprecise category boundaries in a five-point directional scale. As indicated in Table 2, these weights are chosen so that specific disagreement

patterns are successively tolerated in the corresponding estimates of the agreement coefficient. In particular, W_1 represents the test of weights which generate the unweighted kappa measure of perfect agreement proposed by Cohen (1960).

The use of kappa and K_w is restricted to the cases where the number of coders is two. Generalizations are therefore necessary for cases of more than two coders. Fleiss (1971), Light (1971), and later Kraemer (1980) and Davies and Fleiss (1982) extended the concept to more than two coders. Also, in a series of papers, Landis and Koch (1977a, 1977b, 1977c) developed kappa-like measures of agreement from a contingency table point of view which have an intraclass correlation coefficient interpretation.

III. Graphic Presentation of Inter-coder Agreement

Graphics are very effective communication tools. As compared with textual and tabular forms of presentation, relationships between variables are more clearly grasped and easily remembered when portrayed by charts and graphs. Furthermore, charts and graphs can bring out hidden facts and relationships , and can stimulate, as well as aid, analytical thinking and investigation (Fienberg, 1985). Even though the aforementioned agreement statistics can effectively summarize the overall agreement patterns, these statistics cannot reveal the different coding proclivity between coders. Like the proverb -- a picture is worth a thousand words--, if we can construct the agreement chart or graph based on coders' agreement patterns, this may be an excellent visual aid which can reveal many interesting phenomena hidden in the overall agreement statistics such as Pi and kappa. In this section, the "agreement chart" procedure developed by Bangdiwala and Bryan(1987) will be reviewed.

Since all of the above mentioned agreement statistics do not have the graphical representation, Bangdiwala proposed an agreement statistic (B_n) which has a direct graphical interpretation. He used the "mosaic" concept of Hartigan and Kleiner (1984), in which the individual cell entry counts are represented by rectangles or tiles of area proportional to the count. The "mosaic" makes it possible to compare visually sets of counts, suggest hypotheses, check assumptions and interpret results. The concept of utilizing rectangles to represent counts is adapted for producing the agreement chart to assess observer agreement and investigating other relevant hypotheses such as observer bias (Bangdiwala & Bryan, 1987).

Table 2

Description of Hierarchical Weights

Set of weights	Disagreement permitted for agreement statistic.
W_1	No disagreement; requires perfect agreement
W_2	strongly agree (1) with agree (2)
W_3	strongly disagree (5) with disagree (4)
W_4	strongly agree (1) with agree (2), and strongly disagree (5) with disagree (4)

Figure 1 shows the agreement chart based on a hypothetical table with k=5 classification categories. It is constructed as follows: (1) Draw an n*n square, in this case n=231. Thus the big square on Figure 1 represents 231*231matrix; (2) Draw k rectangles of dimensions $(n_{i.})(n_{.i})$, i= 1, 2, ... 5, within n*n square. The rectangles are positioned such that the first rectangle of dimension $(n_{1.})(n_{.1})$ is in the lower left corner of the square, the second rectangle's lower left corner joins the first one's upper right corner, and so on until 5th rectangle is in the upper right corner of th n*n square. These rectangles contain the information
on the marginal totals of the contingency table. The marginal totals of coder A determine the width, and thsoe of coder B decide the length of each rectangle; (3) Blackening diagonal entries representing perfect agreement between coders for respective category. Thus if the two coders agreed exactly, that is, if all off-diagonal cells of two-way table are zeroes, the agreement will consist of diagonal squares rather than rectangles, and they will all be blackened.

Based on the above graphical concepts for two-way contingency table, Bangdiwala proposed an agreement statistic which has the unique graphical interpretation capability. Under independence, the expected value of the diagonal cell entry n_{ii} is the area $n_{i.}n_{.i}$ in the corresponding rectangle, i=1, ..., k. Thus a descriptive measure for agreement between coders that naturally arises from such a chart is the proportion of the total area of the k squares that represents agreement between the two coders. This area is $\sum_i x_{ii}^2$. Dividing by the total area of

the k rectangles, one obtains the statistic

$$B_n = \sum_i x_{ii}^2 / \sum_i x_{i.} x_{.i} \quad \text{---------- (14)}$$

This statistic can be interpreted as the ratio of the sum of the areas of squares of perfect agreement to the sum of the areas of rectangles of marginal totals. Thus, kappa and B_n statistics are quite similar; however, unlike kappa, the lower bound of B_n statistic cannot have negative values.

Similar to the notions of a K_w, the B_n can be generalized to a "weighted B_n" statistic by meaningfully incorporating the information contained in the off-diagonal cell entries. In Figure 1, the areas of disagreement are treated equally (blank) like equal o weights given to off-diagonal disagreement cells in calculating the unweighted kappa statistic; however, "partial agreement" patterns can be visualized in agreement chart using graphic technique. This technique will be reviewed in the next section.

IV. Illustrative example

This section is concerned with the analysis of the fictitious data in Table 3 with primary emphasis given to illustrating the methodology in Sections II and III. Table 3 displays the information on independent classification by two coders (coder A and coder B) on the same subjects using a five-point ordinal scale. The kappa statistic is .213 and the Pi statistic is .197. According to Landis and Koch's nomenclature (1977b, p.165), this represents the poor agreement between coders. The possible maximum value of kappa permitted by marginals of Table 3 is .682. Thus .318 (1-.682) represents the proportion of possibilities of agreement which cannot be achieved as a consequence of different marginals between coders. As Cohen argued, if this value can be used to check the ambiguity of the category boundaries, then there is much room to refine the categories.

Table 3

		Coder A				
	Categories	1	2	3	4	5
	1	38	5	0	1	2
	2	33	11	3	0	5
Coder B	3	10	14	5	6	16
	4	3	7	3	10	18
	5	0	3	7	10	21

The weights described in Eq. 13 were used to calculate K_w. The weight matrix is shown in Table 4. The K_w based on Eq. 9 and 13 is .463. The weighting scheme substantially increases the agreement statistic, not a totally surprising result. As shown in Table 4, the weights near the diagonal entries (perfect agreement) are larger than those of the cells that are further away. Also, Table 3 shows that frequencies of the cells adjacent to diagonals are generally large. In the unweighted kappa statistic, all off-diagonal entries are given zero weights as shown in Table 1. Therefore, it seems that the unweighted kappa statistic, as compared with K_w, is too stringent to use as the intercoder agreement index, especially for ordinal scales, since the statistics does not take into account how close the disagreements are. Of course, for nominal scales, the K_w statistic cannot be used because any mismatch is as good or as bad as any other mismatch.

As has been mentioned, the specific patterns of disagreement between coders can be examined by considering a hierarchy of weights which creates a potentially less stringent measure of agreement by successively combining adjoining categories. The sequence of

Table 4

Weighting Matrix for Five-point Ordinal Scale

	1	2	3	4	5
1	1	.75	.50	.25	0
2	.75	1	.75	.50	.25
3	.50	.75	1	.75	.50
4	.25	.50	.75	1	.75
5	0	.25	.50	.75	1

hierarchical kappa statistics with the four weights given in Table 1 is K_1 = .213, K_2= .319, K_3=.305, and K_4=.440. K_1 is the unweighted kappa statistic with five categories; K_2 with categories 1 and 2 combined, K_3 with categories 4 and 5 combined, and K_4 with categories 1 and 2 combined into one category and categories 4 and 5 combined into another category. Category 3 is not combined with any other categories since in a five-point ordinal scale, the middle category is usually reserved as the neutral one. The results indicate that kappa statistics increase in successive agreement measure. Thus, the coders are exhibiting significant disagreement between adjacent categories (1,2), (4,5), as evidenced by the inflated frequencies in these off-diagonal cells in Table 3. These results suggest that the category boundaries between categories 1 & 2, and categories 4 & 5 are not very distinctive with respect to their usage by two coders. With repect to the relatively relaxed criterion corresponding to the fourth set of weights (W_4), the kappa statistic (K_4=.440) indicates a moderate level of intercoder reliabilty. Thus, three categories instead of five categories may be enough for the data.

It is of some interest to test the homogeneity of marginal distributions between coders. If the distributions are homogeneous, the pooled marginal distribution can be used when calculating PE(expected frequencies). In that case, Pi statistic is preferable since it is computationally simpler than kappa. The test statistic, generated by the formula (A.14)in Appendix 1 in Koch et al. (1977), is Q_c = 54.13 with 4 degree of freedom, which implies that there are significanct ($p < .01$) differences in the response profiles between coder A and

coder B. This results suggest that significant intercoder bias exists between the two coders in their overall usuage of the classification scale.

Figure 1 has essentially the same information as the K_1 statistic. But it shows much more information than K_1 may show. It discloses that coder A has a tendency to code the subjects into two extreme categories. The length of the lower left corner rectangle, which represents the marginals for category 1 by coder A, is about twice as long as the width attributable to coder B. Exactly the same pattern is found in the upper right corner rectangle. Such an information can be extracted by carefully examining the two-way contingency table like Table 3. But the visual relationship portrayed in Figure 1 is more clearly grasped and more easily remembered.

In contrast to Figure 1 which represents the unweighted agreement chart, Figure 2 visualizes the partial agreement. In this figure, the rectangles become progressively lighter in shading the further away they are from the diagonal squares with perfect agreement. In calculating K_w, we consider sets of weights which assign varying degrees of partial credit to the off-diagonal cells depending on the extent of the agreement. Examples of these weights are given in Table 4. Instead of giving numerical values as shown in Table 4 to off-diagonal disagreement cells, different shadings are given to disagreement portions of rectangles in Figure 2. By examining the configuration of rectangles in Figure 2, the reader can easily recognize that the coders are showing substantial disagreement between category (1,2), and (2,3). Figure 3 shows the result of combining categories 1 and 2. K_2 is the numerical counterpart of Figure 3. Figure 4 and Figure 5 are graphic presentation of K_3 and K_4 respectively.

The concept of intercoder bias can be easily visualized in the agreement chart (see Figure 2, for example) by noting the deviation of the path of the marginal totals rectangles (that formed by the positive slope diagonals of the rectangles) from the diagonal straight line of marginal homogeneity. The closer to the diagonal the path is, the closer in agreement the coders' marginal distributions are. Apparently, the path of rectangles in Figure 2 deviates from the straight line of marginal homogeneity. Figure 2 has coder B classifying items more towards the middle categories than coder A, who classifies items more towards extreme categories. This kind of analysis is useful only for ordinal categories, since the rearrangement of the rectangles, possible for nominal scales, make the path of rectangles an arbitrary criterion.

V. Computer Programs

Many of the measures discussed in this paper are practicable by hand-calculator for moderate-sized tables, especially if the item has no more than three categories. The only computations necessary are the observed frequencies of agreement and the expected frequencies of agreement based on the marginal distributions. Also, many of the measures can be calculated using "canned" computer programs. For example, the PROC CATMOD in SAS™ can be used to calculate the kappa and weighted kappa statistics as well as the test statistics for marginal homogeneity (Q_C). A SAS™ program is available from the author for computing kappa, weighted kappa, variance of kappa and weighted kappa based on Fleiss et al.' s(1969) formula, hierarchical kappa based on different weights, and Scott's Pi. This program uses the PROC MATRIX code in SAS™. The SAS/GRAPH™ program generating agreement chart is also available. This agreement chart program was originally developed by Bangdiwala and Brian at University of North Carolina.

Acknowledgement

The author would like to thank Dr. Comstock and Dean Stephens at S.I. Newhouse School of Public Communication for their helpful comments and supports. Many thanks extend to Dr. Bangdiwala and Ms. Bryan at University of North Carolina. They are kind enough to release the "Agreement Chart" SAS/GRAPH™ program to the author. This program provided inspiration and a starting point to this paper.

NOTES

The author can be contacted at the following address:

Namjun Kang
Rm. 24
Communication Research Center
215 University Place
Syracuse University
Syracuse, New York 13244
(315) 423-3363

References

1) Bangdiwala, S.I., Bryan, Hope. Using SAS™ software graphical procedures for the observer agreement chart, *Proceedings of the Twelefth Annual SAS™ Users Group International Conference* , Dallas, U.S.A. 1083-1088, 1987

2) Cichetti, D. V. A new measure of agreement between rank order variables. *Proceedings, 80th Annual Convention*, APA, 1972, 17-18

3) Cohen, J. A. Coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 1960, **20**, 37-46

4) Cohen, J. A. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin*, 1968, **70**, 213-220

5) Davies, M. and Fleiss, J. L. Measuring agreement for multinomial data, *Biometrics*, 1982, **38**, 1047-1051

6) Everitt, B. S. Moments of the statistics k and weighted kappa, *British Journal of Mathematical and Statistical Psychology*, 1968, **21**, 97-103

7) Fienberg, S. Graphical Methods in Statistics, *The American Statistician*, 1979, **33**, 165-184

8) Fleiss, J. L. Measuring nominal scale agreement among many raters, *Psychological Bulletin* **76**, 1971, 378-383

9) Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa, *Psychological Bulletin*, 1969, **72**, 323-327

10) Hartigan, J. A. and Kleiner, B. A mosaic of television ratings, *The American Statistician*, 1984, **38**, 32-35

11) Light, R. J. Measures of response agreement for qualitative data: some generalizations and alternatives, *Psychological Bulletin*, 1971, **76**, 365-377

12) Koch, G, Landis, J. R., Freeman, J., Freeman, D., and Lehnen, R. G. A general methodology for the analysis of experiments with repeated measurement of categorical data, *Biometrics*, 1977, **33**, 133-158

13) Kraemer, H. C. Extension of the kappa coefficient, *Biometrics*, 1980, **36**, 207-216

14) Krippendorff, K. Bivariate agreement coefficient for reliability of data, E. F. Borgatta, ed. *Sociological Methodology*, 1970. Jossey-Bass, San Francisco

15)Krippendorff, K. *Content Analysis: An introduction to its Methodology*, 1980, SAGE Publication Inc., Beverly Hills

16) Landis, R. and Koch, G. A review of statistical methods in the anlaysis of data arising from observer reliabilty studies (Part I), *Statistica Neerlandica*, 1975a, **29**, 101-123

17) Landis, R. and Koch, G. A review of statistical methods in the analysis of data arising from observer reliabilty studies (Part II), *Statistica Neerlandica*, 1975b, **29**, 151-161

18) Landis, r. and Koch, G. The measurement of observer agreement for categorical data, *Biometrics*, 1977a, **33**, 159-174

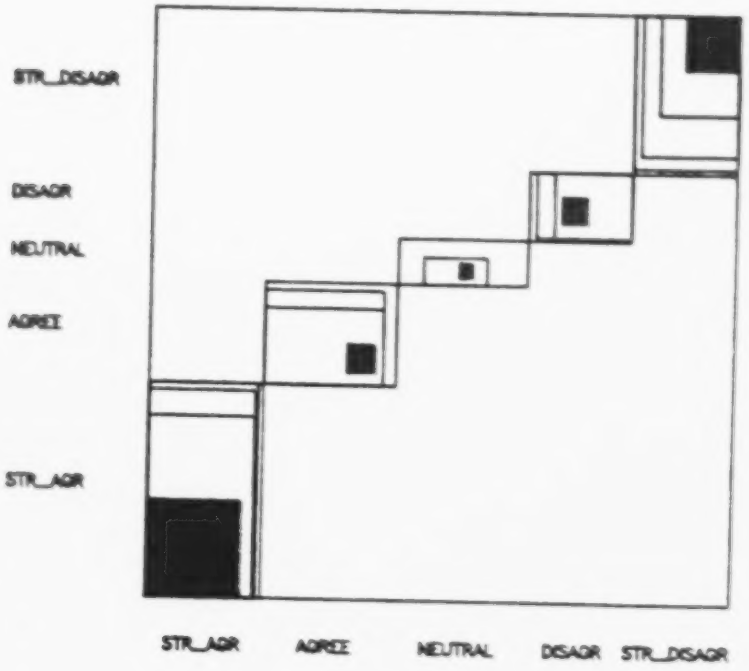
19) Landis, R. and Koch, G. An aplication of hierachical kappa-type statistics in the assessment of majority agreement among multiple observers, *Biometrics*, 1977b, **33**, 363-374

20) Landis, R. and Koch, G. One-way components of variance model for categorical data, *Biometrics*, 1977c, **33**, 671-679

21) Scott, W. A. Reliability of content analysis, *Public Opinion Quarterly*, **19**, 321-325

KAPPA=.213 PI=.197
CODER A

FIGURE 1



CODER B

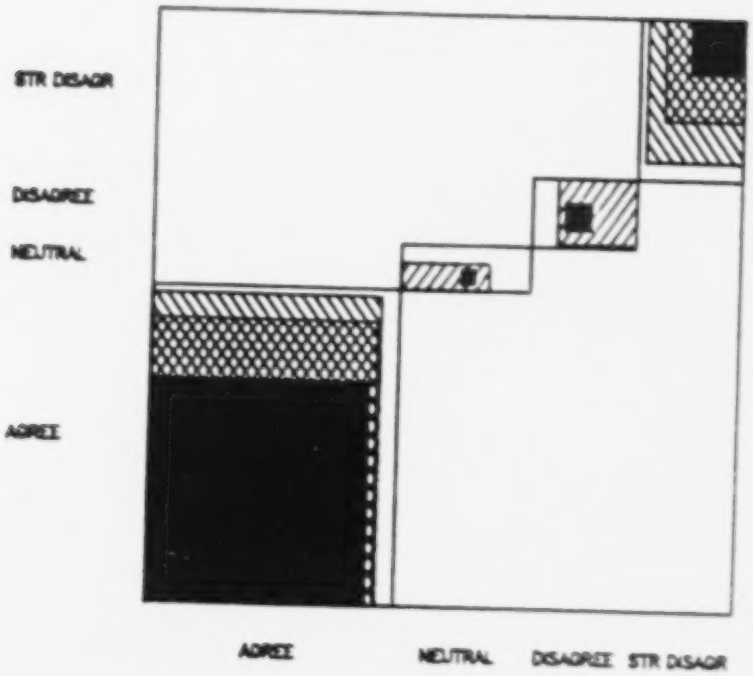
KAPPA=.213 PI=.197
CODER A

FIGURE 2

CODER B

KAPPA=.319 PI=.308
CODER A

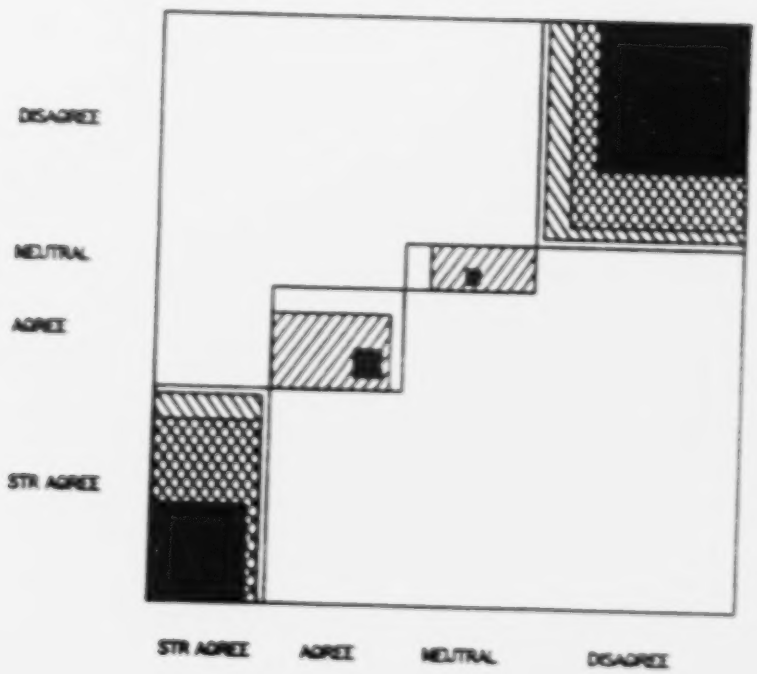
FIGURE 3



CODER B

KAPPA=.304 PI=.292
CODER A

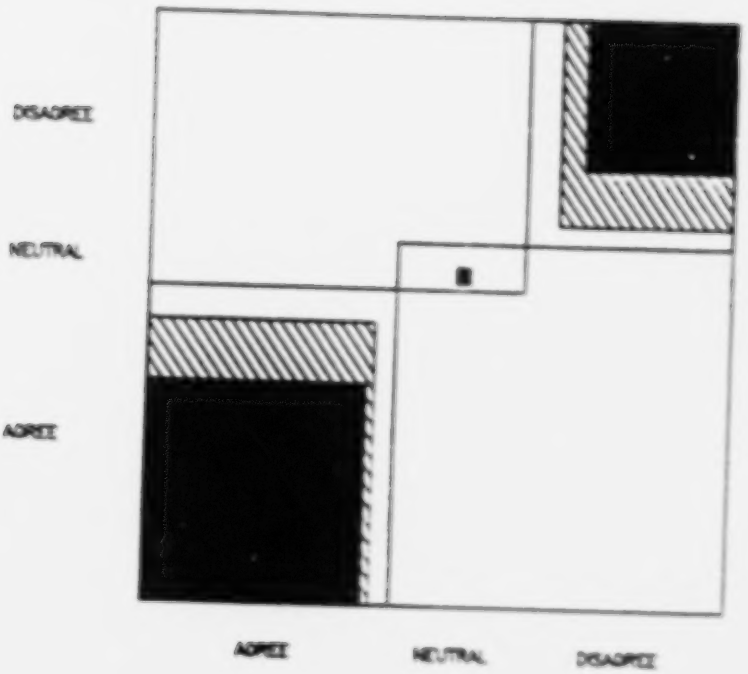
FIGURE 4



CODER B

KAPPA=.440 Pi=.432
CODER A

FIGURE 5



CODER B